

Problem set 1
Econ 121-10
Summer 2020

Part 1 must be submitted (pdf attachment preferred). Part 2 should be answered through Canvas in the quiz section. In the STATA/R exercises, please upload your do file/R-script with comments.

1 Part 1

1.1 Individual project

1. What is the data source you chose? Describe the variables: content, timing, location, etc.
2. Why do you think understanding these variables, and their association, could be interesting to your or to others?
3. Create a histogram and box-and-whiskers plot of your variables
4. Are there outliers? Do you think these outliers could be a problem in the future?
5. Summarize your data: mean, median, 25th percentile, 75th percentile, standard deviation, min, max
6. Are there missing observations in any of the variables? do we know why?
7. Does the data look symmetric? if not, what type of skewness there is?

1.2 Proofs

1. Let x_1, x_2, \dots, x_n be observed values of a variable x . Let \bar{x} be the average and let s be the standard deviation of these observations, so $z_i = \frac{x_i - \bar{x}}{s}$ is the i^{th} value of x expressed in standard units. Show that the average \bar{z} expressed in standard units is zero and the standard deviation is one.
2. Consider the following model of linear statistical association:

$$y_i = b_1 x_i + u_i$$

Given observations $(y_i; x_i)$ for $i = 1, \dots, n$; derive an expression for the slope coefficient that minimizes the average squared difference between the observed values of y and the values predicted by this regression

1.3 STATA/R

Import the Auto data.

Stata: `webuse auto.dta, clear`

R: `install.packages("ISLR") library(ISLR)`

1. Find the mean and the median of the variable weight.
2. Draw a histogram to show the distribution of the variable weight.
3. Draw a quantile function of the variable weight and find the 25%, 50%, 75% percentiles.
4. Draw a box plot for weight. Are there any outliers?
5. Draw a scatterplot for price and weight and find the correlation between these 2 variables

6. Draw the scatterplot again with a best-fit line. Does the relation look linear?
7. Regress price on weight. What is the coefficient on the x variable ($\hat{\beta}_1$)? What is the intercept in each equation ($\hat{\beta}_0$)? How can we interpret the coefficient on weight?
8. Include variable length on the regression. What are the coefficients of a regression of price on length and weight ($\hat{\beta}_1$ and $\hat{\beta}_2$)? What is the intercept ($\hat{\beta}_0$)?
9. Create a correlation matrix for the three variables

2 Part 2

Please submit in canvas.